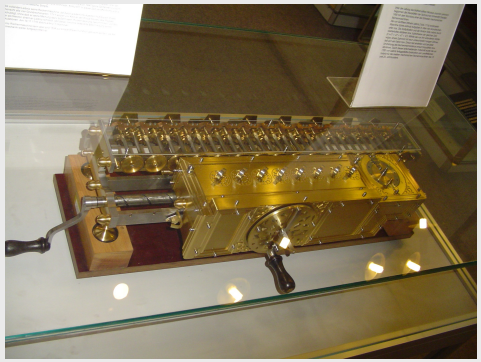
LRZ’s recent Altix 4700 installation

High-Performance Computing at Leibniz Computing Centre Munich, Germany

Iris Christadler, LRZ High Performance Systems Department

HPC in Bavaria

www.lrz.de



The Leibniz Computing Centre

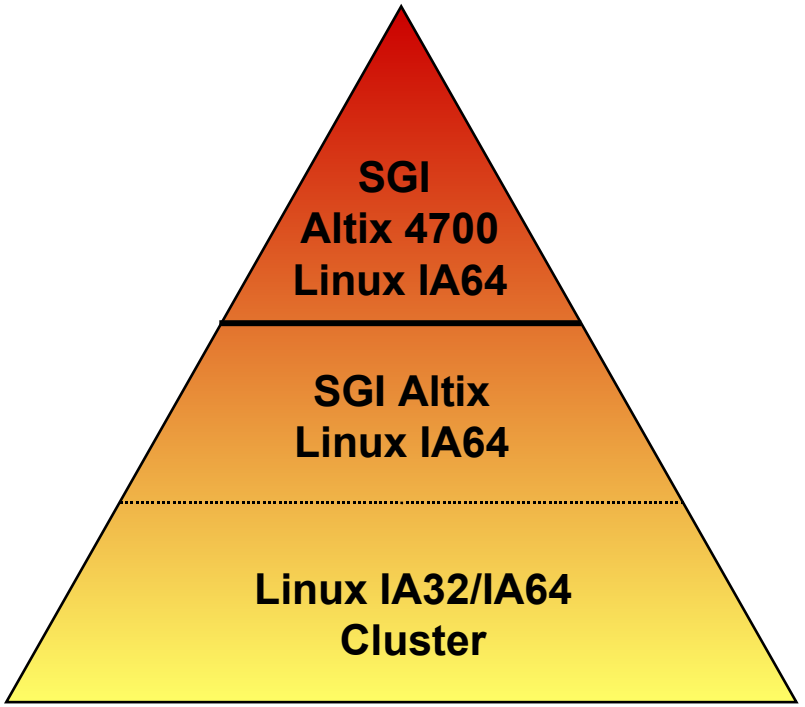
- Named after Gottfried Wilhelm Leibniz (1646 – 1716)
- Computing Centre for all Munich Universities
- HPC Centre for all bavarian researchers (128-way SGI Altix+Linux Cluster, 3.2 TFlops in total)
- National HPC Centre for Germany (9728 cores SGI Altix 4700, 62 TFlops peak perf.)
- Member of
  - Gauss Centre of Supercomputing
  - Munich Scientific Computing Network (MCSC)
  - DEISA
  - D-Grid
  - AstroGrid

Moving to Itanium (since 2004)

National Capability

Regional Capability & Capacity

Local Capacity



- Switch to Linux for regional and local systems
  - Itanium2, Pentium, and EM64T (AMD Opteron +Intel Nocona)
  - High synergy and acceptance by users
  - Reduction in cost for support, staff and licenses
- National/Capability Scale
  - Still focus on high sustainable application performance
  - Outstanding offer of an Itanium Linux system by SGI

www.lrz.de

LRZ's HPC Systems 2007

System			#Cores	Peak (TFlops)	Memory (TByte)	Disks (TByte)
National System	SGI Altix 4700 (Phase2)		9728	62.3	39	660
Porting and Tests	SGI Altix 4700		256	1.6	1	10
Linux Cluster	IA32		138	0.8	0.2	50
	EM64T (Xeon, Opteron)		50	0.3	0.1	
	IA64 Itanium	2-way	134	0.8	0.8	
		4-way	68	0.4	0.1	
		8-way	16	0.1	0.024	
		SGI ALtix 128-way SMP	128	0.8	0.5	
		Subtotal	348	2.1	1.4	
	Total		534	3.2	1.8	50

www.lrz.de

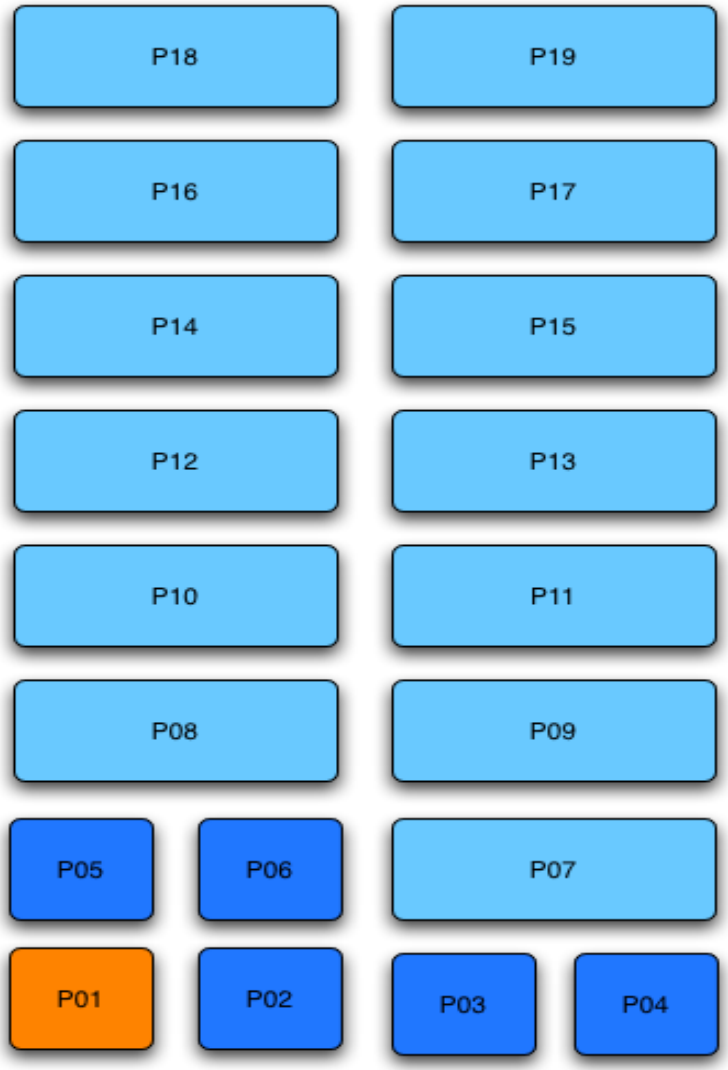
SGI Altix 4700

Overall characteristics for both installation phases:	Phase 1 (until 03/2007)	Phase 2 (since 04/2007)
Total number of cores	4096	9728
Peak Performance (entire system)	26.3 TFlops	62.3 TFlops
Linpack Performance	24.5 TFlops	**> 56.2 TFlops**
LRZ-Benchmark Performance	8.2 TFlops	> 16.2 TFlops
Size of memory (entire system)	17.5 TByte	39 TByte
Direct Attached Disks	300 TByte	600 TByte
Network Attached Disks	40 TByte	60 TByte
Processor type	Madison	Montecito Dual Core
Clock rate	1.6 GHz	1.6 GHz
L3 Cache (per core)	6 MByte	9 MByte
Memory per core	4 GByte	4 GByte
Clock rate of frontside bus (FSB)	533 MHz	533 MHz
Peak bandwidth to local memory	8.5 GByte/s per core	8.5 GByte/s shared (between 2 or 4 cores)

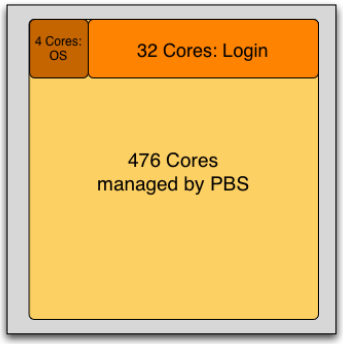
SGI Altix 4700 Phase 2: Partition Layout

www.lrz.de

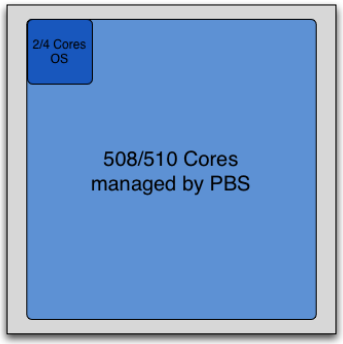
- 512 Cores per partition
- 19 Partitions managed by PBSPro
- 13 High-Bandwidth partitions (light blue)
  - 2 cores share memory channel
  - 2 cores reserved for OS
- 6 High-Density Partitions (aqua + orange)
  - 4 cores share memory channel
  - 4 cores reserved for OS
  - 32 cores reserved for login in partition1
- Interactive jobs for performance measurements via PBS
- Advanced reservation possible for LRZ staff



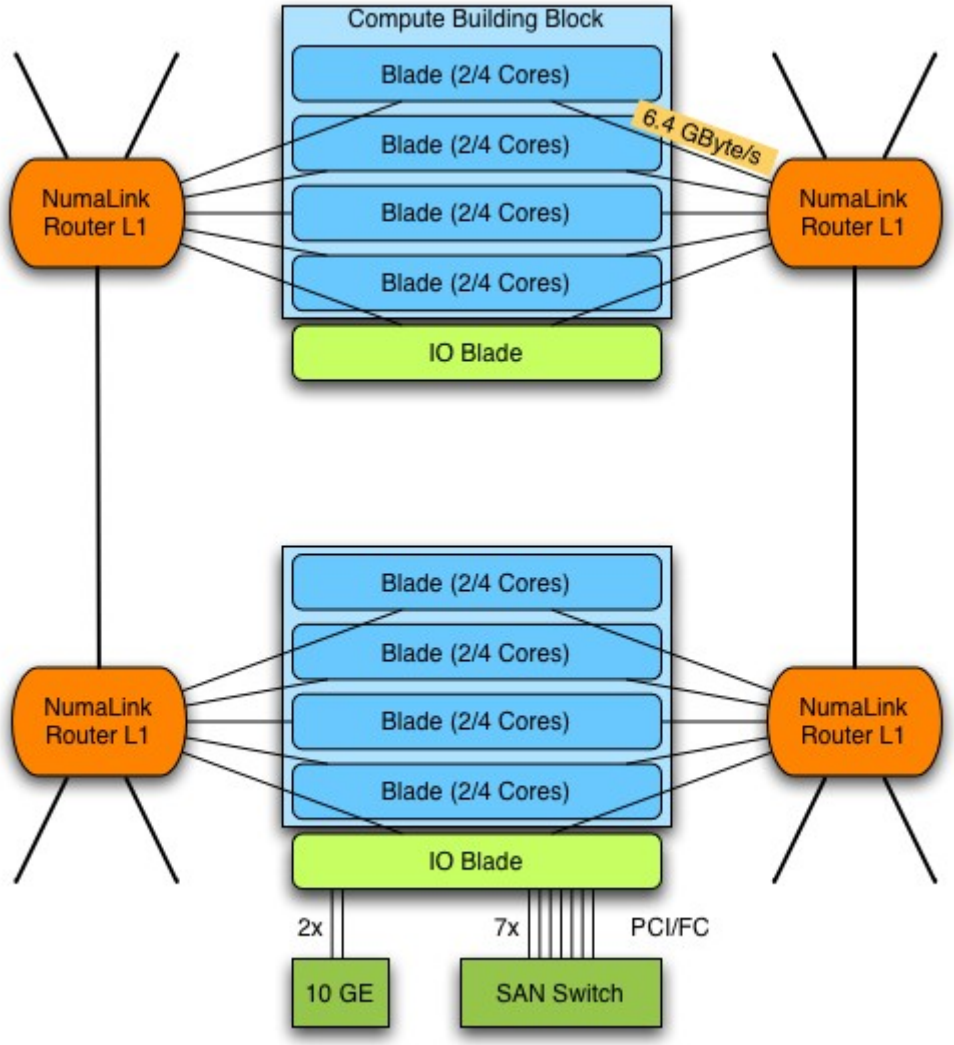
Login Partition



Regular Batch Partition



SGI Altix 4700 Phase 2: NUMA Link Building Block



Benchmark Suite

Application Benchmark Results

Program	#Cores used for one image	System Aggregated Performance [GFlops]	Weight
RINF	8	1646,28	0,09
FFT	8	5305,68	0,06
SIPSOLVER	8	4501,50	0,06
ZHEEVD	2	33886,21	0,05
DMRG	1	18985,07	0,05
LASER	1	15359,49	0,04
LINPACK	9728	56514,33	0,05
BEST	1024	10197,47	0,13
BQCD	1024	7960,00	0,11
CACTUS	384	17629,24	0,09
HEPFP	64	22381,49	0,09
MGLET	512	8297,94	0,11
NWCHEM	64	38180,22	0,07
Total		**16207,88**	**1,00**

From Single to Dual Core

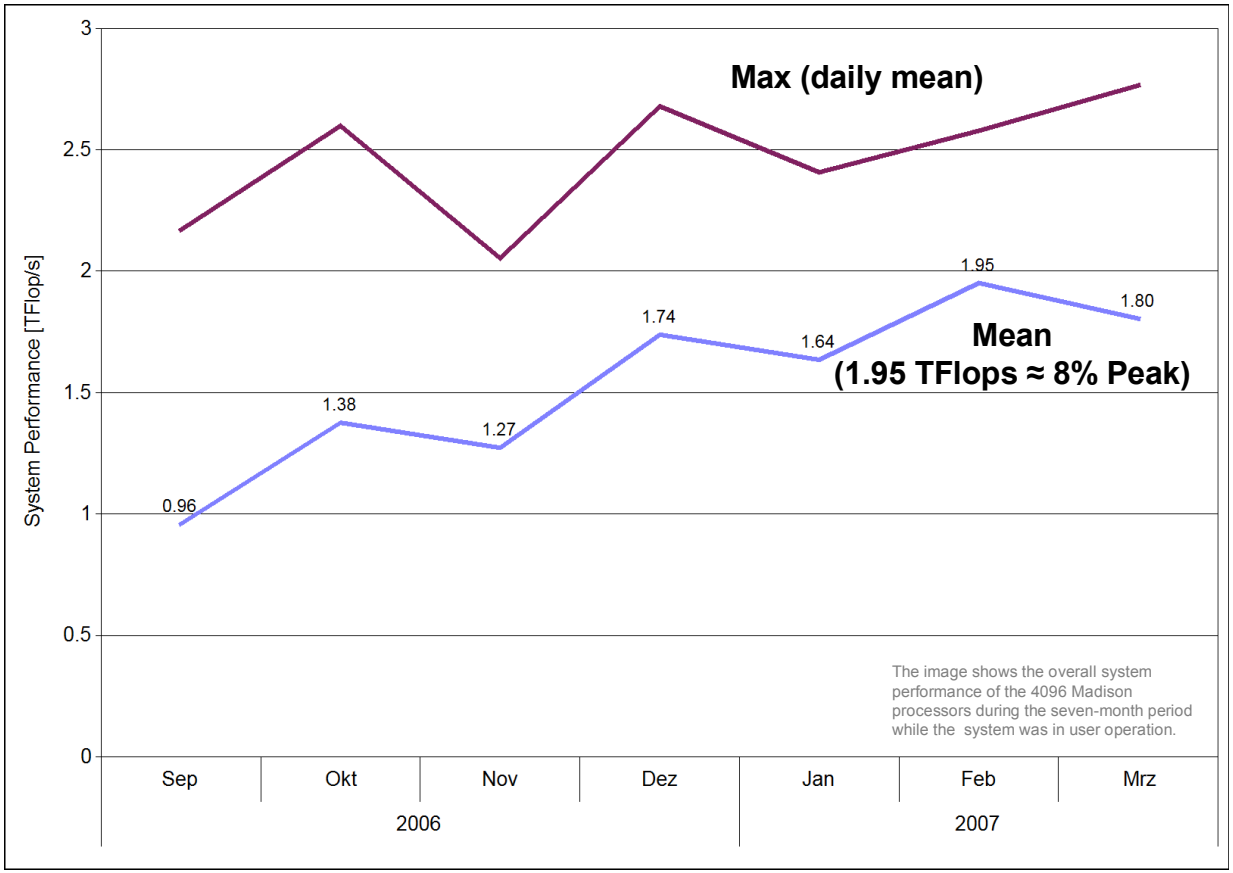
Sustained System Performance (Phase 1)

- Sample ~40 Itanium PMEs (5 minute intervals, system-wide)
- Store results into an SQL-Database
- Add user info to the database

► monitor system performance

► draw conclusions on various performance aspects

(see http://www.lrz-muenchen.de/wir/berichte/TB/LRZ-Bericht-2006-06.pdf)

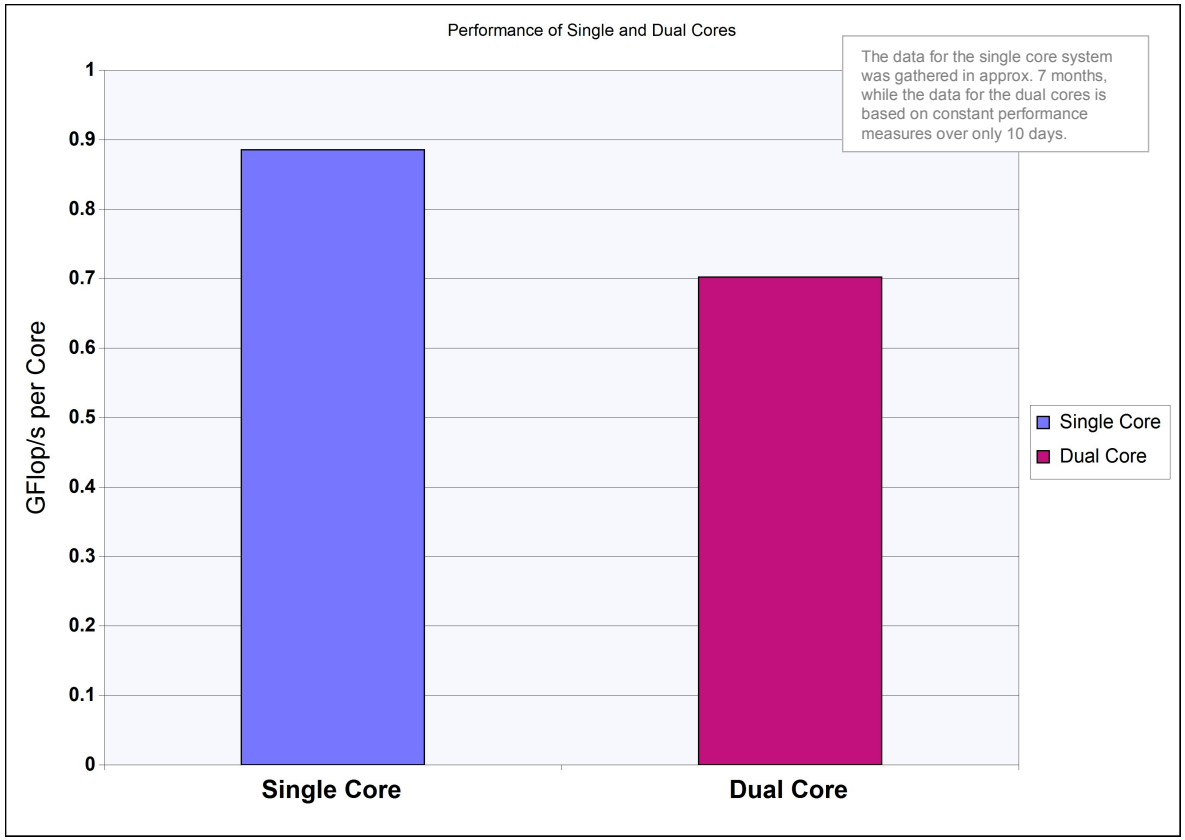


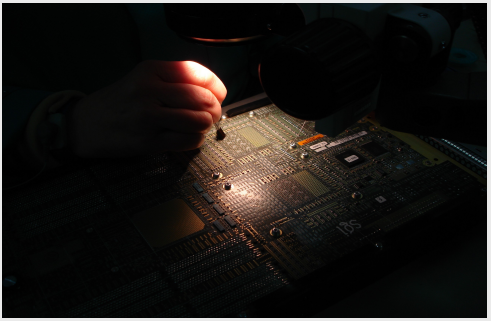
www.lrz.de

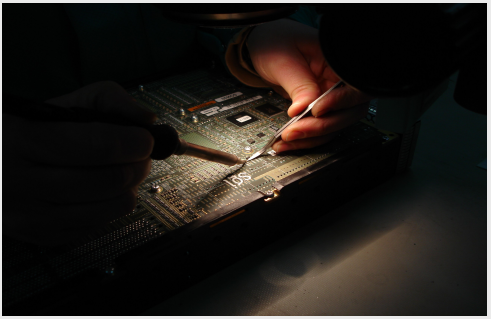
SGI Altix 4700: Switching from Single to Dual Core

- Switch from Madison to Montecito Dual Core (March 2007)
- Increase of L3 cache per core (6MB → 9MB) to reduce memory pressure
- Two cores now share one memory channel

► **Per-core performance dropped by only ~ 20%**

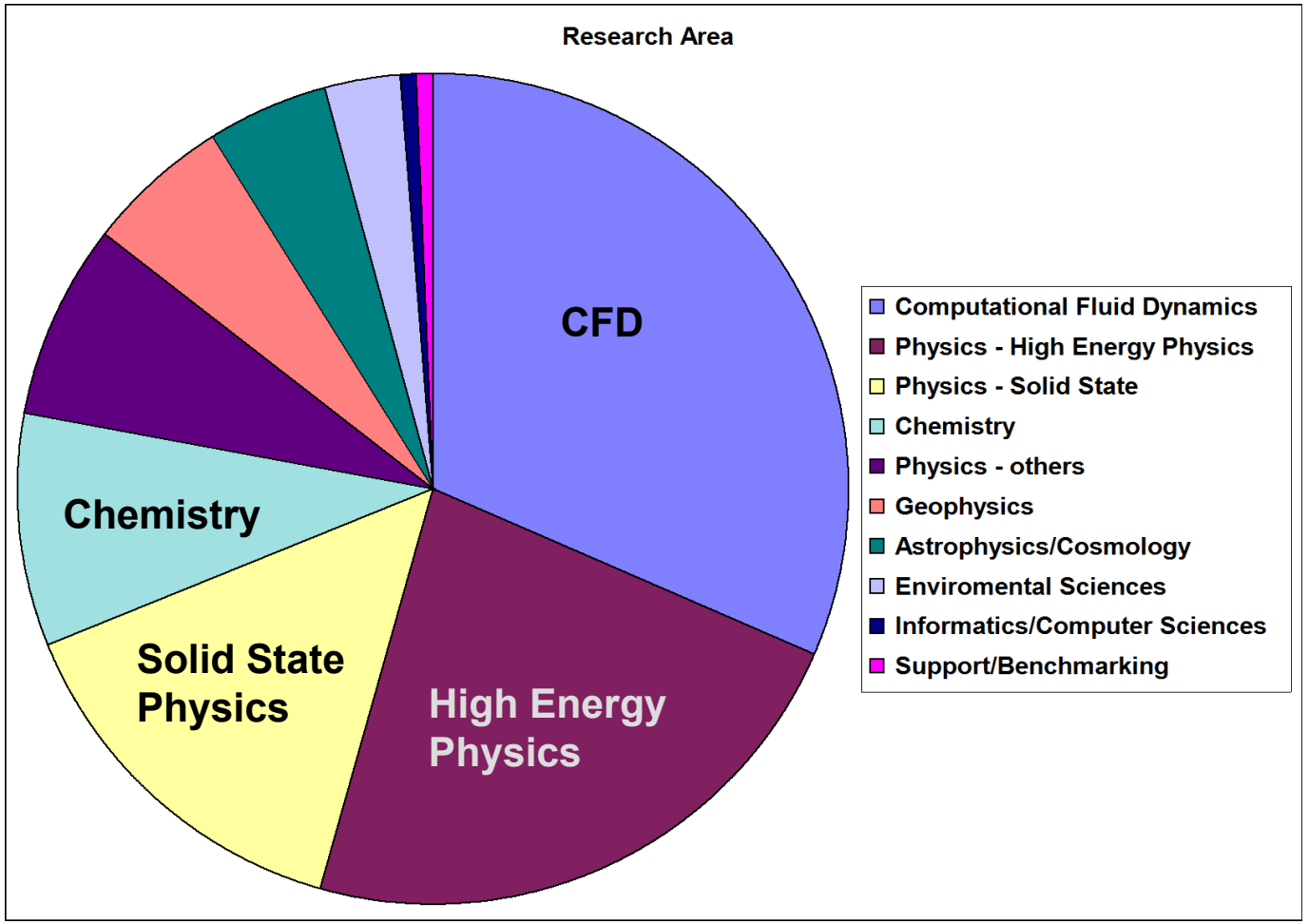






www.lrz.de

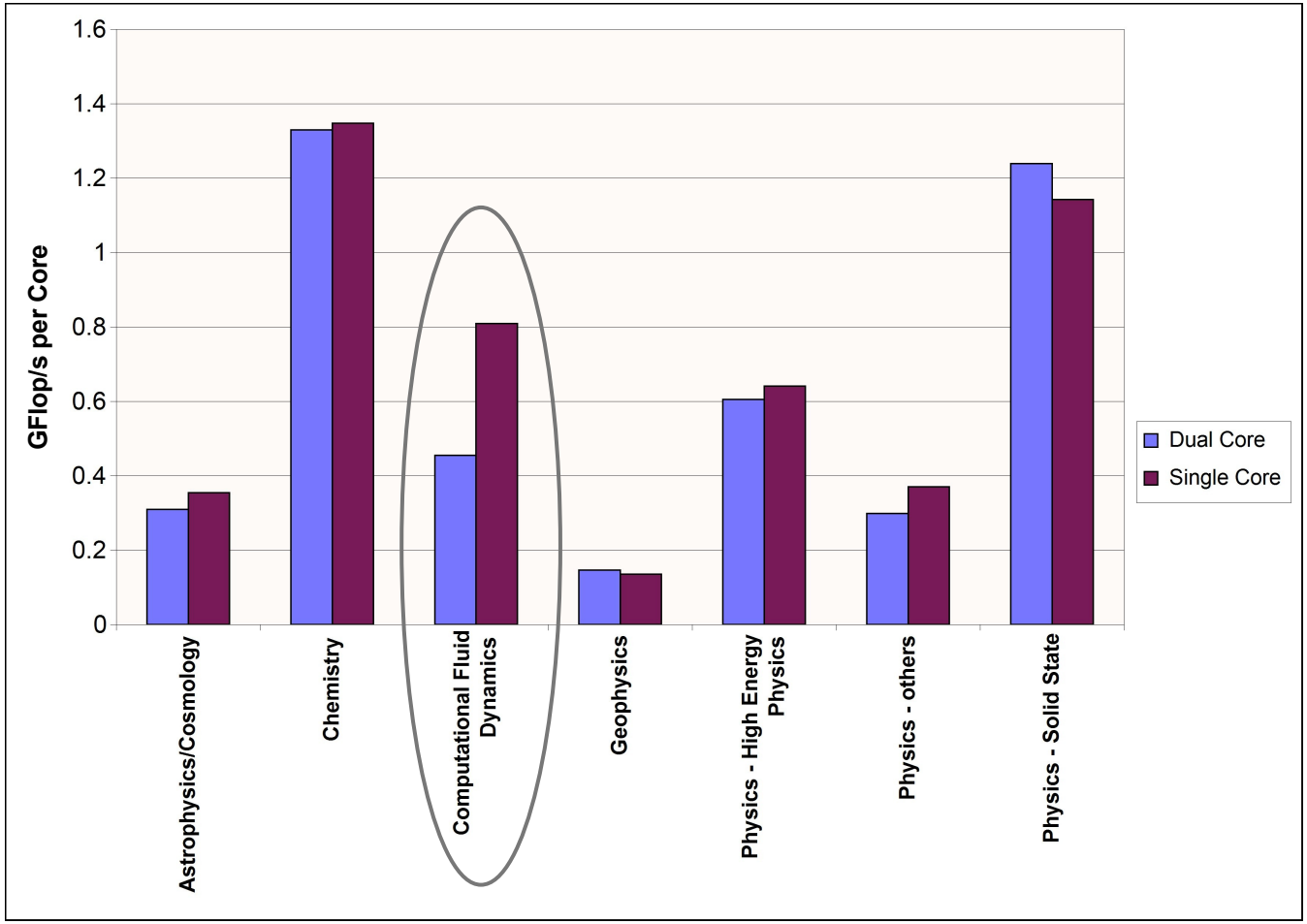
Our Users (in percentage of cycle usage)

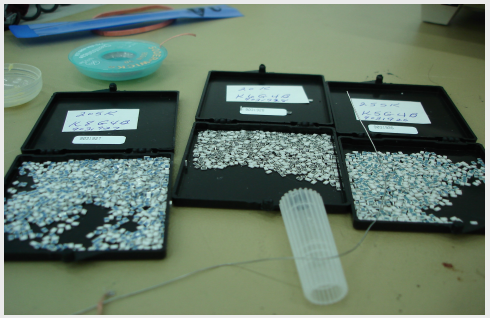


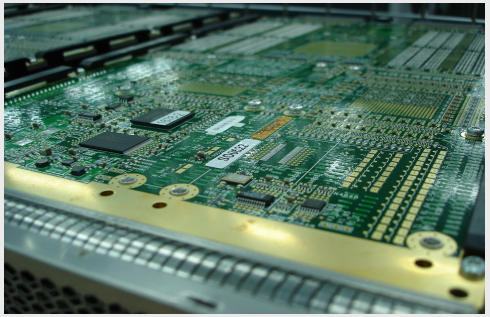
www.lrz.de

Switching from single to dual-core: Performance Data for different research areas

- Most research areas maintain their per-core performance
- Only the very memory intensive CFD codes suffer from the switch





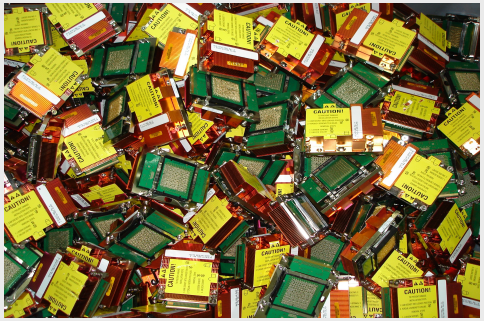


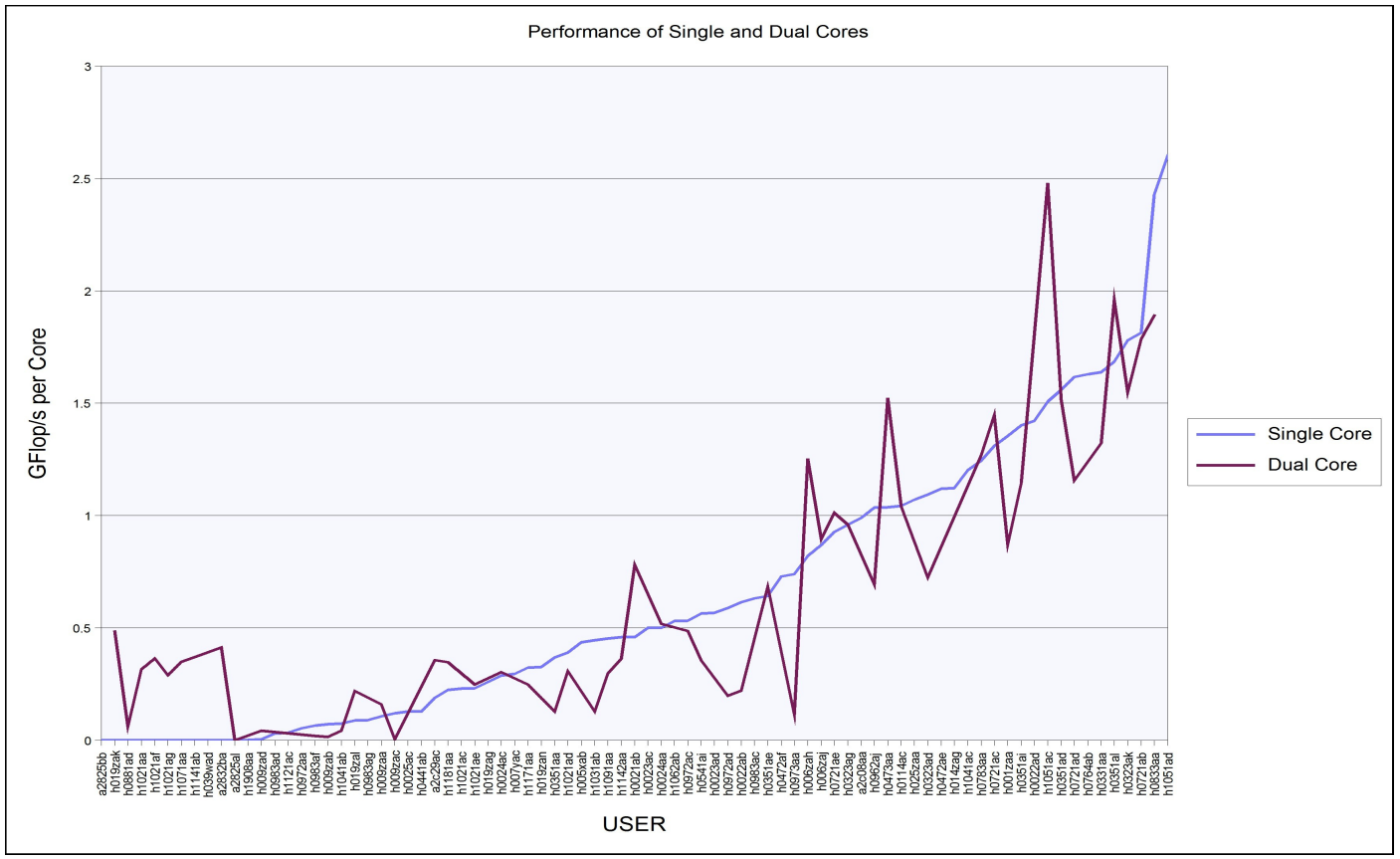
www.lrz.de

Switching from single to dual-core: Performance resolved by User

The image shows the first per-user impact of the dual-core update.

- Highly optimized codes still perform well

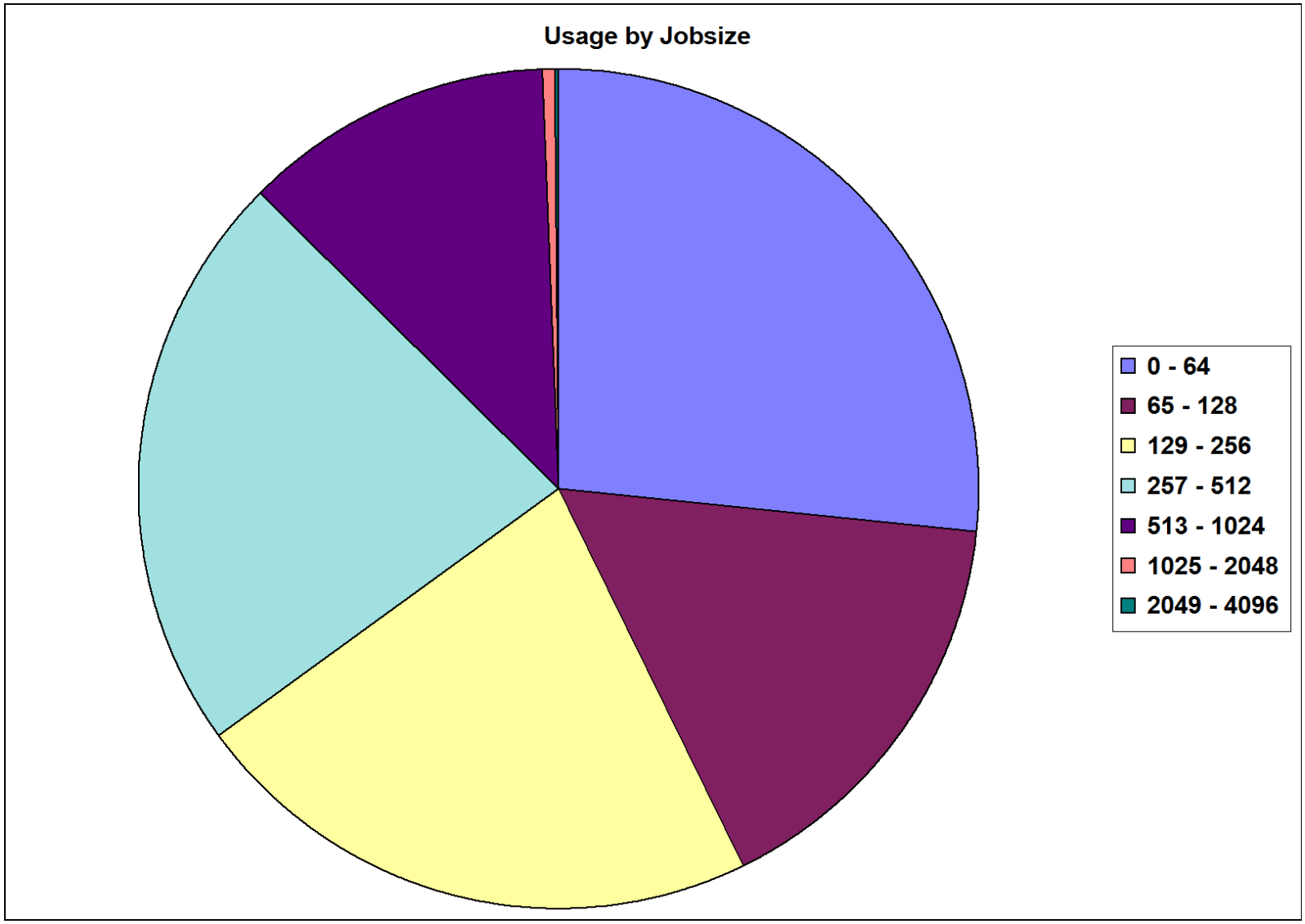




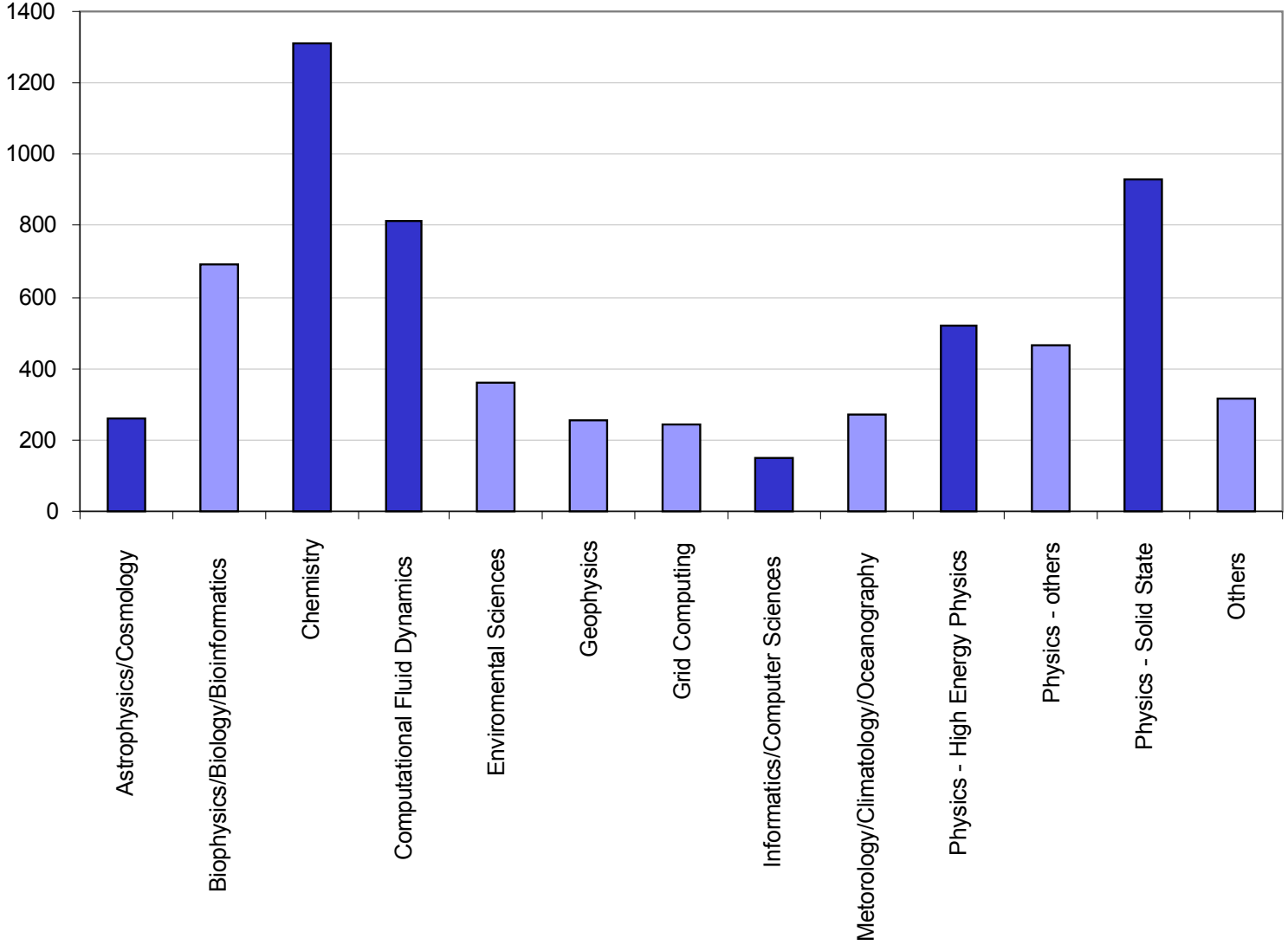
Performance Results

of 4096 Madison Cores

Statistics:
Usage by Job size

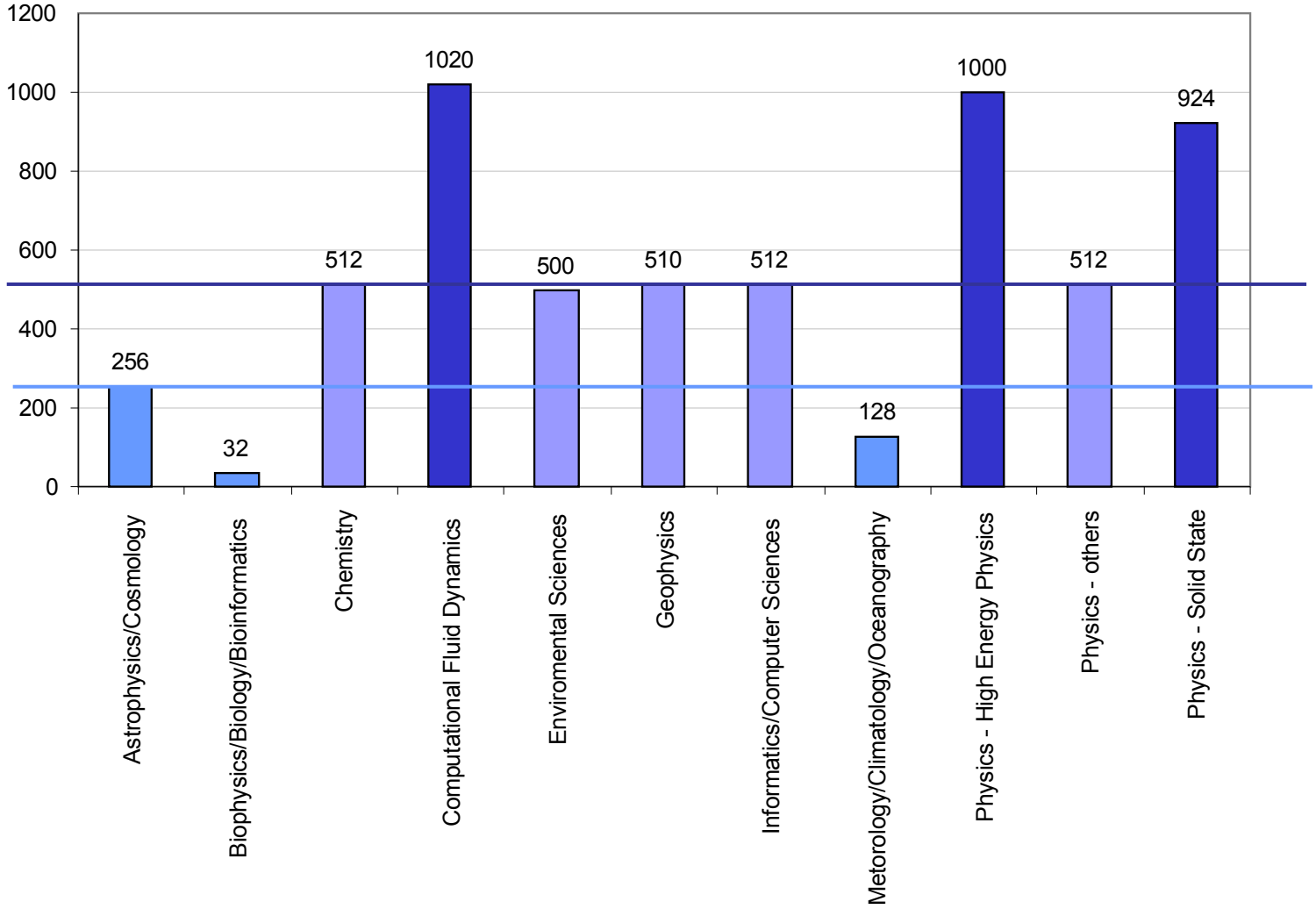


Statistics:
Mean Value of MFlops per Core for different Research Areas

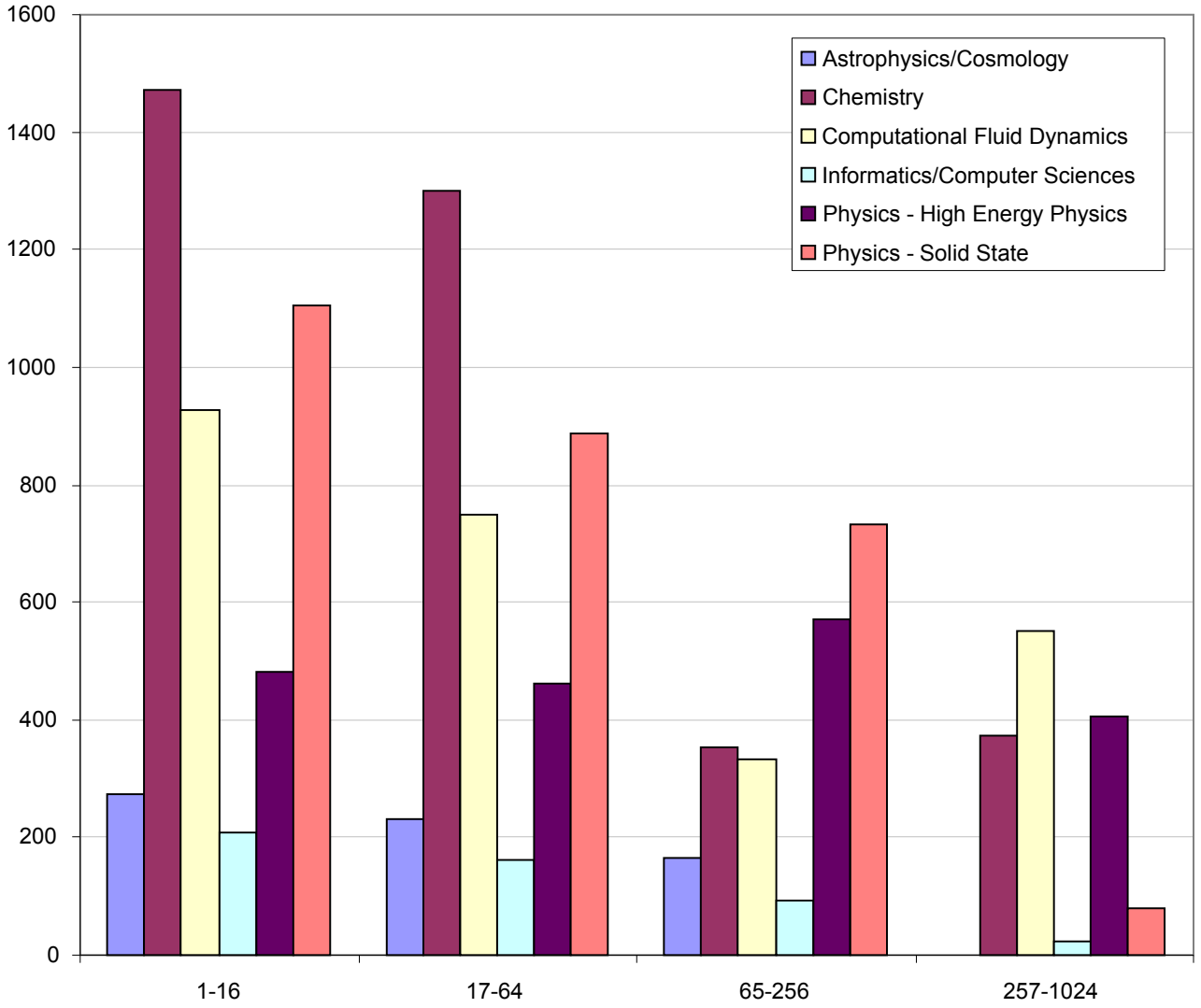


www.lrz.de

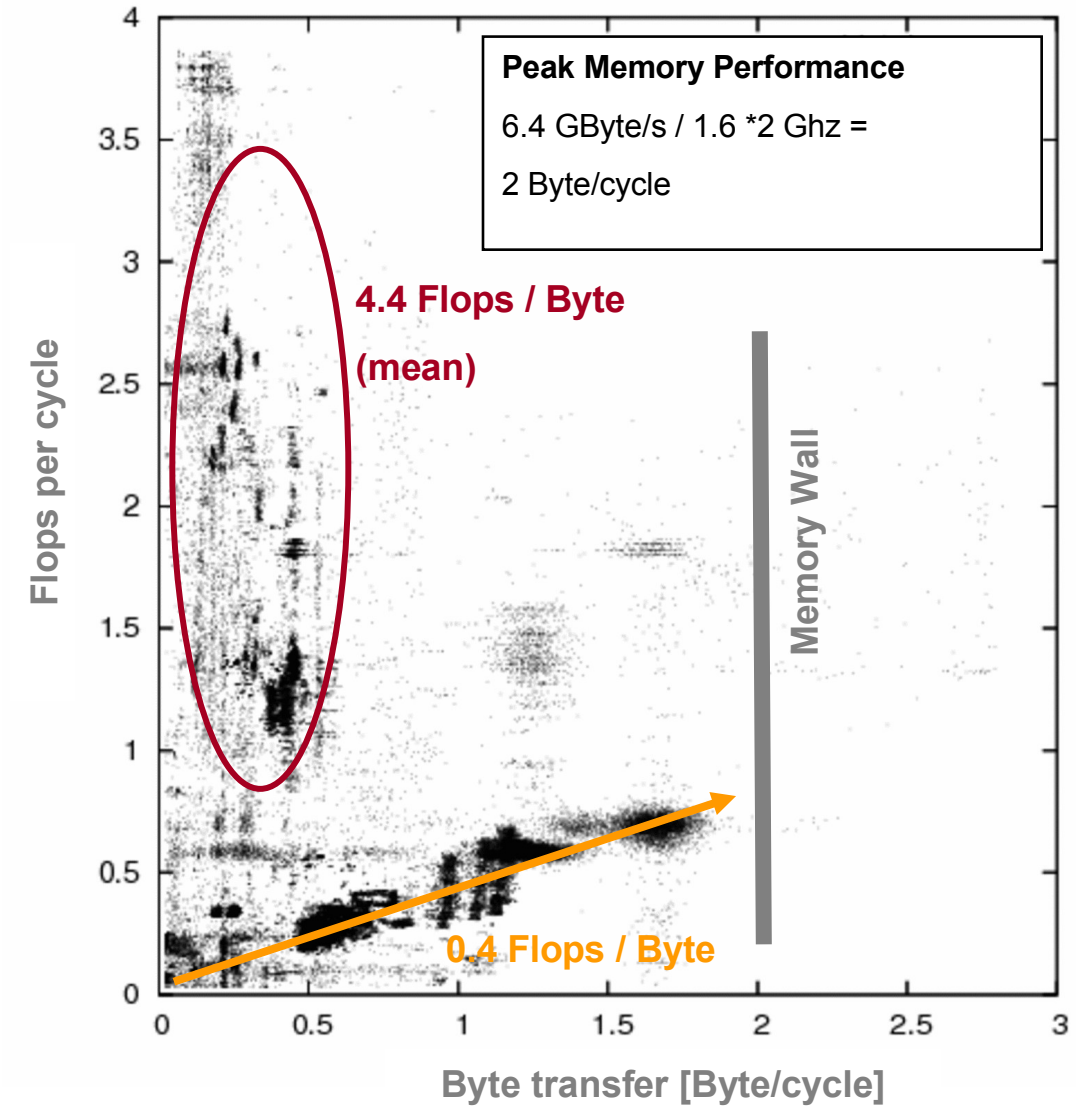
Statistics:
Maximum of used cores for different Research Areas



Statistics:
Mean MFlops per Core vs. Number of Cores



Are we indeed hitting the memory wall?



Detailed Analysis of applications that ran on our small 128-way SGI Altix 3700 Bx2 (120000 samples).

Overall:

1.4 Flops / Byte